Published in *Autism Research*

Investigating diagnostic bias in autism spectrum conditions:

An item response theory analysis of sex bias in the AQ-10

Aja Louise Murray[1,2*], Carrie Allison[3], Paula L. Smith[3]

Simon Baron-Cohen[3], Tom Booth[2], & Bonnie Auyeung[2, 3]

[1]Violence Research Centre, Institute of Criminology, University of Cambridge, UK

[2]Department of Psychology, University of Edinburgh, School of Philosophy, Psychology and Language Sciences, UK

[3]Autism Research Centre, Department of Psychiatry, University of Cambridge, UK

* Corresponding author: Institute of Criminology, Sidgwick Avenue, Cambridge, CB3 9DA.

Email: am2367@cam.ac.uk

Lay Abstract

Previous research has suggested that there are some important differences between males and females with autism spectrum conditions (ASC) with, for example, females being better able to hide or compensate for their difficulties. ASC is also more often associated with males than with females and this tends to be reflected both in perceptions of the conditions and in the kinds of symptoms and behaviours included in assessments used to identify and diagnose them. Together, these factors may make it more difficult to identify ASCs in females. In this study, we tested whether a recommended screening tool for ASC, the AQ-10, worked the same way in males and females and was free of bias against females. We found individual items that differed across the sexes, but these cancelled out when considering the test as a whole. While this supports the continued use of the AQ-10 as a brief screen for ASC, it suggests a need to be cautious about interpreting responses to individual items.

Abstract

Diagnostic bias is a concern in autism spectrum conditions (ASC) where prevalence and presentation differ by sex. To ensure that females with ASC are not under-identified, it is important that ASC screening tools do not systematically underestimate autistic traits in females relative to males. We evaluated whether the AQ-10, a brief screen for ASC recommended by the National Institute of Clinical Excellence (NICE) in cases of suspected ASC, exhibits such a bias. Using an item response theory approach, we evaluated differential item functioning (DIF) and differential test functioning (DTF). We found that although individual items showed some sex bias, these biases at times favoured males and at other times favoured females. Thus, at the level of test scores the item-level biases cancelled out to give an unbiased overall score. Results support the continued use of the AQ-10 sum score in its current form; however, suggest that caution should be exercised when interpreting responses to individual items. The nature of the item level biases could serve as a guide for future research into how ASC affects males and females differently.

Keywords: sex differences; autism; screening; autism spectrum quotient; AQ -10; item response theory; differential item functioning; differential test functioning

Autism spectrum condition (ASC) are characterised by difficulties in social communication and interaction alongside restricted interests and repetitive behaviour (American Psychiatric Association, 2013). The impairments associated with ASC can have a significant impact on functioning and well-being. Individuals with ASC may experience difficulties living independently, forming friendships, and gaining employment (Eaves & Ho, 2008) and adults with ASC are at higher risk of experiencing suicidal thoughts than people in the general population (Cassidy et al., 2014).

Facilitating access to relevant services and resources for individuals with ASC and their families relies on accurately identifying the condition. Many individuals who would meet the criteria for a diagnosis of ASC are currently not known to relevant ASC services (Baron-Cohen et al., 2009; Kim et al. 2011) and investing effort in identifying these individuals could bring much needed assistance to those who need it. Full diagnostic assessment for ASC is a time and resource intensive process and places a potentially significant burden on the individual being assessed. It would be unsustainable and undesirable to submit individuals to this process without evidence for the appropriateness of such assessment, especially given that the prevalence of ASC is around only 1% in the population (Baird et al., 2006; Baron-Cohen et al., 2009). To this end, brief screening tools for ASC have been developed to quickly assess suspected cases of ASC in order to identify individuals who may have undiagnosed autism, whilst filtering out those for whom a full diagnostic assessment is unlikely to be appropriate.

The AQ-10 was developed as a brief screen for ASC for use with adults with average or above average intellectual functioning (Allison, Auyeung & Baron-Cohen, 2012). It is a 10 item self-report measure that can be administered by frontline clinicians and social care

professionals. The National Institute for Clinical Excellence (Guideline 142; NICE, 2012) recommends administering the AQ-10 to individuals suspected of having an ASC and offering full diagnostic assessment to those who score above the cut-off point of 6 (based on binary item scores). Studies evaluating the performance of the AQ-10 against the criterion of clinical diagnosis have suggested that it performs well in identifying individuals with an ASC (Allison et al., 2012; Booth, Murray, McKenzie, Kuenssberg, O'Donnell & Burnett, 2014). However, an important outstanding issue in the use of the AQ-10 and other screens for ASC is the possibility that their use contributes to a diagnostic bias; specifically, the under-identification of females with ASC.

It has been noted that when a clinical condition varies in its prevalence and/or presentation across males and females, there is the potential for diagnostic bias – the under- or over-diagnosis of one sex relative to the other (Rutter, Caspi & Moffitt, 2003). In ASC, sex ratios vary dependent on subtype, level of functioning, and population, but there is little doubt that males outnumber females, with overall sex ratios in the region of 2:1 to 4:1 (Fombonne, 2009; Baron-Cohen et al., 2011). Furthermore, sex differences in ASC have been reported at the genetic, physiological and behavioural level (Lai et al 2013; Lai et al., 2015). Several authors have suggested that, reflected in this male preponderance and perhaps because of possible female compensatory mechanisms, females with ASC are less likely to be successfully identified than males with similar levels of impairment (Krieser & White, 2014; Lai et al. 2015; Baron-Cohen, Lombardo, Auyeung & Chakrabarti, 2011). For example, Russell, Steer and Golding (2011) found that, when controlling for ASC trait severity, males were more likely to receive a diagnosis of ASC than females. Similarly, among those who receive a diagnosis of ASC, females are on average older at the point of diagnosis (Begeer et al. 2013; Giarelli et al. 2010; Rutherford et al. 2016; Shattuck et al. 2009). These observations are consistent with the evidence that in order to receive a diagnosis of ASC, females with

ASC must display more severe symptoms or present with additional problems relative to males (Dworzinski, Ronald, Bolton & Happé 2012; Zwaigenbaum et al. 2012). Factors that contribute to females being under-recognised may include greater social motivation to try to fit in or to try to 'camouflage' difficulties, the use of another girl in the peer-group as a model for social learning and possibly better language and/or imitation skills (e.g. Lai , Lombardo, Auyeung, Chakrabrti & Baron-Cohen, 2015)

If there are diagnostic biases in ASC, females with ASC may be less likely to gain access to services and resources from which they can benefit. Furthermore, because much research into ASC relies on clinically diagnosed samples, any diagnostic biases that lead to an under-representation of females are translated into non-representative research samples and, in turn, biased substantive conclusions. The need for females to display more severe symptoms to receive a diagnosis of ASC may, for example, partly explain the observation that in clinically diagnosed samples, females can sometimes show more severe ASC traits and comorbid psychopathology than males (Carter, Black, Tewani, Connolly, Kadlec & Tager-Flusberg, 2007; Dworzynski et al. 2012; Hartley & Sikora, 2009). Therefore, ensuring that the process of identifying ASC is not biased against females is an important goal for ensuring fairness of diagnostic procedures in a clinical context and accuracy of substantive conclusion in a research context.

Given the potential for bias in identifying females with ASC, and the use of screening tools in the referral of individuals for full clinical ASC assessment, one important question is whether these screening tools display a gender bias. Although, owing to a lack of research, there has been little direct evidence to suggest that screening tools are biased in this way, the possibility has been raised and is consistent with the generally 'male-focussed' process of development and evaluation of assessments for ASC (Kreiser & White, 2014). That is, it has been argued that the inclusion of symptom and behavioural indicators in ASC assessments

has been influenced by the perception that ASC is primarily a 'male condition'. The problem is compounded by the fact that the validation of ASC assessments has utilised predominantly male samples. It is therefore a concern that items from commonly used and recommended ASC screens may function differently according to whether the respondent is male or female. Females with ASC may, for example, fail to endorse some items because they refer to more typically male manifestations. In this case, the items will be expected to show differential item functioning (DIF).

An item can be said to show differential item functioning (DIF) by sex if a male and female of the same level of ASC traits have different probabilities of endorsing that item (Magis, Béland, Tuerlinekx & De Boeck, 2010). This logic can be extended to differential test functioning (DTF) where the expected total score on a test differs for a male and female of the same level of ASC trait. It is possible for DIF to occur without DTF if there are some items biased in favour of females that balance out items biased in favour of males. However, if there are biases that do not cancel out, the test can systematically under- or over- estimate the ASC levels of females relative to males, or vice versa. Clearly this is undesirable in screening for ASC.

Identifying DIF can also provide new insights or highlight undetected differences in the presentation of ASC between males and females. Although they did not formally assess DIF, Kopp and Gillberg (2011) found some evidence that males and females with ASC have different likelihoods of endorsing certain items of the Autism Spectrum Screening Questionnaire –Revised Extended version (ASSQ-REV). Regarding DSM-IV (APA, 2000), ASC diagnostic indicator of having friends that are appropriate to developmental age, boys with ASC were more likely to endorse an item indicating that they lacked best friends, whereas girls were more likely to endorse an item that indicated that they interact mainly

with younger children. The study by Kopp and Gillberg (2011), therefore, highlighted one way in which social deficits in ASC may manifest differently in males and females.

Given the importance of DIF and DTF for fair screening practices, for ensuring that substantive conclusions regarding sex differences are not skewed by diagnostic bias, and the potential for it to yield new insights into sex differences, it was our aim in this study to evaluate DIF and DTF in the AQ-10.

Method

Participants

We used archival data comprising a combined sample of individuals with a clinical diagnosis of ASC and control individuals. Our rationale for doing so was twofold. First, at the point of use of a screening tool, it is not known whether an individual meets the criteria for a diagnosis of ASC, therefore, it is more representative of how the AQ-10 is used in practice to use a combined sample that is agnostic to ASC status. Second, previous research has highlighted that restricting analyses to data from only clinically diagnosed or control individuals risks biasing statistical results through range restriction, assuming that ASC traits are on a continuum (Murray, McKenzie, Kuenssberg & O'Donnell, 2014). Previous research has suggested that such a continuum is captured by the AQ: the larger instrument from which the AQ-10 is derived (Murray, Booth, McKenzie & Kuenssberg, 2015).

There were 557 females and 680 males who reported a diagnosis of ASC included in our analyses. Participants were recruited online via the volunteer database of the Autism Research Centre (www.autismresearchcentre.com). The majority of the sample reported having a clinical diagnosis of Asperger Syndrome (AS, N=998) or High Functioning Autism (HFA, N=158). Other reported diagnoses included pervasive developmental disorder (PDD),

PDD not otherwise specified (PDD-NOS), autism, atypical autism, autism spectrum condition (ASC) and autism spectrum disorder (ASD). Individuals were included if they reported having a clinical diagnosis of ASD and provided details of their diagnosis. Diagnoses were pre-existing and not administered by the research team. Individuals were not selected for, or excluded from, the sample based on AQ or AQ-10 scores. The clinically diagnosed sub-sample had a mean AQ-10 score of 8 (SD=1.97) and a mean age of 35.02 years (SD=13.10). The majority was of White European (Caucasian) ethnicity.

There were 4,462 female and 2,894 male controls included in our analyses. All controls were recruited online via the volunteer database at www.cambridgepsychology.com, and none reported having a first-degree relative with a diagnosis of ASC. As expected, the control sub-sample had a lower mean AQ-10 score of 2.86 (SD=2.02) but were similar in other respects. The difference in age between the two sub-samples was statistically significant (t (1,705.9) =7.9, p<.001) but this partly reflected the very large sample size as the actual mean age of 31.82 (SD=13.50) of the control sub-sample was similar to that of the clinically diagnosed sub-sample. Given the very small association between age and AQ-10 scores (r (8,590) =.05, p<.001) this difference in age was not judged to be problematic. Like the clinically diagnosed sub-sample, the control sub-sample was also of majority White European origin. The total sample size for the current analysis was 8,593 (female = 5,019; 58.4%).

Measures

AQ-10

The AQ-10 is a brief 10 item self-report screen for ASC (Allison et al., 2012). Items ask the participants to rate the extent to which they agree with a statement about their behavioural preferences or tendencies by selecting one of four response options 'Strongly

Agree', 'Definitely Agree', 'Slightly Disagree' and 'Definitely Disagree'. Four of the items are phrased such that selecting 'Strongly Agree' indicates high levels of autistic traits and six are phrased in the opposite direction where selecting 'Strongly Disagree' indicates high levels of autistic traits. The AQ-10 is then scored on a dichotomous response format by assigning both 'Strongly Agree' and 'Agree' responses a numerical value of 0 (or 1 depending on the direction in which the item is phrased) and the 'Disagree' and 'Strongly Disagree' responses a numerical value of 1 (or 0). The individual item scores are then summed to give a score out of 10. Previous research has suggested that scoring above 6 is an indicator that an individual may have an ASC (Allison et al., 2012). Although, dichotomising scores reduces the precision of the instrument; this scoring system is currently preferred for two reasons: first, it is simpler and more practical for the frontline professionals who use the tool in practice and second, validation studies to date have been based on this scoring system (Allison et al. 2012; Booth et al. 2014; Murray, Booth et al., 2015). As it is this scoring system that is used in clinical practice, we adopted it for the current study in order to reflect the screening process as it actually occurs, acknowledging the possibility that this screening process may not be optimal from the perspective of maximising precision of trait-level estimates.

The scale was developed from the full 50 item Autism Spectrum Quotient (AQ; Baron-Cohen, Wheelright, Skinner, Martin & Clubley, 2001) by selecting the 2 items that showed the best discrimination between individuals with a diagnosis of ASC and controls within each of the 5 subscales of the AQ. Thus, two items each were selected from the 'Attention to Detail', 'Attention Switching', 'Communication', 'Imagination,' and 'Social' subscales of the full AQ. After selecting these items, Allison et al. (2012) assessed the ability of the AQ-10 to successfully categorise individuals as case versus non-case using Receiver Operator Characteristic (ROC) curve analysis. The scale performed well, yielding an area under the curve (AUC) of .95 and a sensitivity and specificity at the cut-off point of 6 of .88

and .91 respectively. In an independent sample using a similar methodology, Booth et al. (2014) found that the AQ-10 could discriminate between individuals with and without a clinical diagnosis of ASC cases and controls with sensitivity and specificity of .80 and .87 respectively at the suggested cut-point of 6. In this study, all participants were administered the full 50-item AQ but only the 10 items of the AQ-10 were selected for analysis. Although all 50 items of the AQ were available, we focussed on the AQ-10 rather than the full AQ because it is the former that is recommended for use as a screen for ASD in clinical practice due to its brevity and ease of administration.

Statistical Procedure

Preliminary analyses

We assessed item bias by sex using an item response theory (IRT) approach that assumes that items measure a single underlying construct (unidimensionality). We began by evaluating whether the 10 items of the AQ-10 formed a reasonable approximation to a unidimensional scale in males and females separately. We used several methods to evaluate this: parallel analysis with principal components analysis (PA-PCA), the minimum average partial (MAP) test and examination of a scree plot. We also ensured that a single factor model provided good fit in a confirmatory factor analysis. To account for the binary response format, we used weighted least squares means and variances (WLSMV) estimation. Scaling and identification were achieved by fixing the latent variable variance to 1. Models were estimated in *Mplus 6.11* (Muthén & Muthén, 2010). We judged the models to be good fitting if RMSEA was <.08 and were TLI and CFI>.95 (Hu & Bentler, 1999; Yu, 2002).

Differential Item Functioning

We assessed Differential item functioning (DIF) using an IRT approach. Item responses were modelled using 2 parameter logistic model (2PL) which uses the following model to represent the probability of endorsing an item in terms of a logistic function of the difference between the trait level of the individual and the location of the item on the trait continuum:

$$P_j(\theta_i) = \frac{1}{1 + \exp[-\mathrm{a}_j(\theta_i + b_j)]}, \quad (1)$$

where θ_i is the latent trait level for individual i, and a_j and b_j are the discrimination and location parameters for item j respectively. The advantage of using the 2PL is that both uniform and non-uniform DIF can be identified (Magis et al., 2010). Uniform bias occurs when only the b_j parameter differs by group and non-uniform bias occurs when the a_j parameter differs by group.

Uniform bias suggests that the degree to which males (or females) are more likely to endorse an item than females (or males) of a comparable underlying level of autistic traits is the same across the entire range of the AQ-10. That is, both the direction and the size of item bias is uniform across autistic trait levels. There are two types of non-uniform bias. First, ordinal uniform bias is when the degree of bias in an item varies across autistic trait levels but it is always the same group that is more likely to endorse the item given their latent trait level. That is, in ordinal non-uniform bias, only the size and not the direction of bias varies across latent trait levels. This could happen if, for example, an item was only biased for individuals who were high in autistic traits or of the degree of bias in an item became larger as trait levels neared the clinical range. Disordinal non-uniform bias is when not only the degree but also the direction of the bias depends on autistic trait levels; for example, when females are more likely to endorse an item at low autistic trait levels but males are more likely to endorse it at

high autistic trait levels. It is important to test for non-uniform bias particularly in instruments such as the AQ-10 that employ cut-points to select individuals because non-uniform bias could result in serious bias around the cut-point even if the test as a whole seems to show little overall bias. Furthermore, the differences in the degree and direction of bias across different trait levels may provide some insights into how autistic traits manifest differently in males and females depending on autistic trait levels, rather than assuming that sex differences are the same across all levels.

DIF can be visualised by plotting item characteristic curves (ICCs), which show the relation between latent trait level and the probability of endorsing an item. Figure 1 shows a hypothetical example of uniform bias in the 2PL. The two lines represent the ICCs for an item as administered to two different groups, such as males and females. Here the ICCs are parallel and differ only in their location but not in their slope. Figure 2 shows a hypothetical example of an item showing (ordinal) non-uniform DIF. Here the ICCs are non-parallel: the slope for one of the groups has a steeper gradient. Only when there is no (uniform or non-uniform) DIF will the item characteristic curves (ICCs) for an item will be identical for males and females.

IRT models were, unless otherwise stated, estimated using the mirt package in R statistical software using expectation maximisation-based estimation (Chalmers, 2012; R Core Team, 2014). To test for DIF, one group is chosen as the reference group (here males) and the other is the focal group (here females). The model in eq. 1 is estimated in both groups on a common metric obtained by using items identified as non-DIF as anchors fixed equal across groups. This allows a direct comparison of the parameters for males and females. However, as the presence of DIF items can mask or promote the spurious detection of DIF in other items, an item purification process has been recommended as first step (Kim & Cohen, 1995). First, all the items are tested for DIF under an initial assumption of no DIF. Based on

this, those items that are identified as showing DIF are removed from the set and the process repeated without them. This is repeated until a set of items with no DIF has been obtained. This set is then used as the basis for transforming the parameters of the focal group on to the same scale as the reference group. After this final transformation, DIF is evaluated for all items, including those that were previously excluded from the set. To identify an initial set of non-DIF items we used the difR package in R statistical software (Magis et al., 2010), which automates this procedure.

Using a set of item as anchors identified as non-DIF in a first step, we transformed the male and female parameters to be on the same scale and conducted our main tests of DIF and DTF. This was achieved by estimating a multi-group 2PL model with the a and *b* parameters for the non-DIF items fixed equal across males and females. The statistical significance of DIF in the remaining items was evaluated by comparison of the fit of a model with and without the a and *b* parameters fixed equal across groups. Using this method, the presence of DIF was indicated when a chi-square difference test suggested a significant deterioration in fit with the addition of these constraints. However, as trivially small differences between models can easily be significant in such large sample sizes, we also examined information theoretic criteria. Smaller (more negative) values of Bayesian Information Criterion (BIC), sample size adjusted BIC (saBIC) and Aikaike Information Criterion (AIC), indicate a better fitting model. When the difference in BIC between two models was >10, this was taken to suggest strong enough evidence to consider an item to show DIF of potential practical significance. Raftery (1995) classifies a BIC difference >10 as 'very strong' evidence in favour of the better fitting model, with values between 6 and 10 representing 'strong' evidence; values between 2 and 6 representing 'positive' evidence; and values between 0 and 2 representing 'weak' evidence. We chose this stricter criterion because we were interested in DIF that was likely to have an effect that mattered in practice. ICCs for items identified as

DIF based on this criterion were plotted in order to provide further insights into the nature of the DIF.

Differential test functioning

Differential test functioning (DTF) was assessed by examining the differences in expected total scores for males and females given the same ASC trait levels. First, test characteristic curves (TCCs) for each males and females were obtained by summing the 10 item characteristic curves for each group. For the female group we used the item characteristic curves derived after transformation of item parameters to the same scale as the male group. We evaluated the overall bias in the AQ-10 by inspection of the similarity of the TCCs. We focused on scores and latent trait values around the cut-off point of 6. This allowed us to evaluate whether it was likely that males scoring around this cut-point actually had lower trait levels than females scoring around this cut-point.

We also conducted several formal tests of DTF. To assess overall bias in the AQ-10 we computed the signed (sDTF) and unsigned DTF (uDTF) using the method described by Chalmers, Counsell and Flora (2016). sDTF is a measure of the average directional bias and the uDTF is a measure of the average absolute bias, irrespective of which group it favours. As the AQ-10 has a maximum total score of 10, sDTF for this test can range from -10 (completely biased in favour of females) to 10 (completely biased in favour of males). The uDTF for the AQ-10 has a possible range of 0 (no bias) to 10. The sDTF and uDTF will be identical if the TCCs for males and females do not cross at any point. We evaluated the statistical significance of sDTF using the method described in Chalmers et al. (2016). In brief, standard errors and confidence intervals are obtained using an imputation-based method where the standard error of sDTF and uDTF is estimated using the standard deviation of the estimated sDTF and uDTF across the imputed datasets. Significance tests are not currently

available for uDTF because its lower bound is zero, thus complicating the ability to test the null hypothesis that uDTF=0 in the population; therefore, we report 95% confidence intervals for both uDTF and sDTFs but significance tests for sDTF only.

As the most important question regarding bias in the AQ-10 is whether it is biased around its cut-point of 6, we also computed latent trait values in males and females that corresponded to this cut-point and evaluated DTF and its statistical significance at these points on the latent trait continuum.

Results

Preliminary Analyses

Proportions of item endorsement and mean AQ-10 total scores for males and females are provided in Table 1. Item numbers refer to the item numbers from the original full length AQ and all items are coded in the direction of endorsing an item indicating a higher level of ASC. The AQ-10 total scores suggest that in the current sample, males showed higher average autistic trait scores when cases and controls were combined; however, the DIF and DTF methodology does not require that the two groups have equal trait distributions. Despite differing in trait levels, however, the pattern of item endorsement was similar across males and females. For example, item 5 (noticing small sounds) was the most endorsed item while item 20 (reading fictional character intentions) was the least endorsed item for both sexes.

Unidimensionality test

PA-PCA indicated 1 dimension for females and 2 for males; MAP indicated 1 dimension in both sexes and examination of scree plots indicated 1 strong general dimension for both sexes (for females the first eigenvalue= 5.2 while the second eigenvalue=1.00; for males the first eigenvalue=4.8 and second eigenvalue=1.1). Fit statistics for the single group

CFAs are provided in Table 2. These indicated that a unidimensional model was a good fit according to conventional model fit criteria (Hu & Bentler, 1999; Yu, 2002). We also examined 2-factor exploratory solutions to see if they yielded substantively meaningful second factors; however, in both males and females, these yielded one general factor and one minor factor defined by a single item pair. Overall, these tests suggested that the data were sufficiently unidimensional to allow us to proceed with the IRT analyses assuming unidimensionality.

DIF analysis

The initial iterative item purification procedure identified items AQ5 and AQ20 as non-DIF, therefore, these items were used as anchors to place the female parameters on the same scale as the male parameters. The 2PL model parameter estimates for males and females are provided in Table 3. The DIF tests are also provided in this table. Based on the chi-square different test there was statistically significant DIF in items 28, 32, 37, 41 and 45. For items 37, 41 and 45 the BIC difference suggested that the DIF was not practically significant (>10) although for item 37, the BIC difference of 9.7 suggested it was close to practically significant.. The male and female ICCs for the items showing evidence of practically significant DIF are provided in Figures 3 and 4. For AQ28 the bias favoured females, i.e. for the same latent trait level, females were more likely to endorse the item than males. For AQ32, the bias was in the opposite direction and favoured males.

DTF analysis

The TCCs for males and females are provided in Figure 5. Visual inspection of male and females TCCs suggested that they were very similar. The sDTF value was -0.02 (p=.33; 95% CI=-0.08 to 0.03) and the uDTF value was 0.0 (95% CI= 0.02 to 0.12). sDTF for the AQ-10 was not statistically significant overall (p=.33).

DTF at the AQ-10 cut-point

The horizontal line in Figure 5 represents the test cut-off score of 6. The vertical line represents the latent trait level at which males crossed this threshold (= 1.07). This value was very similar to the latent trait level at which females crossed this threshold (= 1.04). At the latent trait value of 1.07 corresponding to the cut-point of 6 in males, sDTF was 0.09 and not statistically significant (p=.25). At the latent trait value of 1.04, the value at which female expected scores were at the cut-point of 6, sDTF was also 0.09 and not statistically significant (p=.24).

Discussion

In the current study we evaluated whether males and females of equivalent autistic trait levels were liable to score differently on the items of the AQ-10, indicating differential item functioning (DIF). We also evaluated whether the expected test scores on the AQ-10 differed for males and females of the same ASC trait levels (differential test functioning; DTF). Five items showed statistically significant DIF, but only two of these could be considered practically significant. More importantly, the direction of bias was not consistent and the biases in favour of males and females balanced out at the level of the test score. This meant that overall there was no appreciable DTF either around the cut-off point of 6 or across the rest of the latent trait distribution. The lack of overall bias in test scores generally supports the use of the AQ-10 as a brief screen for ASC in both males and females, albeit with the caveat that it is unbiased only through bias cancellation and not through a lack of item-level bias.

Whether or not the fact that some items showed DIF is a serious problem is a subject for debate. Some methodologists have recommended that during scale development and evaluation process, items should be assessed for DIF by key sub-groups (here sex), with those

items showing DIF being candidates for exclusion (Sass, 2011). However, this must be weighed against the need for screening assessments to include items that are best able to discriminate between individuals with and without ASC. Given that the AQ-10 items were selected to maximise discrimination between individuals with ASC and controls (Allison et al. 2012) and that the results of the current study suggest that the test as a whole is unbiased with respect to sex, we would argue that the AQ-10 achieves a good balance between these considerations. Thus, we recommend that it continues to be used in its current form when considering the total score. Of course, it is important to acknowledge that biases in individual items remain; they merely cancel one another at the level of the whole test. Thus, caution is due when interpreting responses to individual items because it cannot be said that endorsement of certain items has the same meaning for males and females in terms of their latent trait levels.

Another possibility would be to explicitly acknowledge sex differences in ASC and use information about sex differential item discrimination and difficulty to select the optimal items for screening to maximise accuracy in males and females separately. That is, to develop a separate female AQ-10 and male AQ-10. This would move the focus on to maximising diagnostic accuracy overall and explicitly acknowledge the idea that ASC is likely to manifest differently in males and females. Such an idea would represent a logical extension of the idea that 'female ASC' may require special attention with respect to timely identification and support needs (e.g. Hallady et al., 2015).

The DIF effects were observed were in items AQ28: 'I usually concentrate more on the whole picture, rather than the small details' and AQ32: 'I find it easy to do more than one thing at once'. While it is not possible to be certain what the cause if the DIF in these cases is, we can suggest some speculative explanations that could help inform future research.

Consider first item AQ28 from the Attention to Detail domain referring to global versus local processing style. It showed a pattern of DIF whereby it was: 1) slightly more discriminating in males and 2) more likely to be endorsed by females given ASC level. That is, females were more likely to report attending to small details versus the whole picture than males of the same ASC trait level. This is consistent with the evidence that males have a tendency towards adopting global/holistic strategies while females tend to opt for more local/piecewise strategies on a range of tasks (e.g. see Pletzer, 2014). The DIF result implies that researchers should consider controlling for ASC trait level – or other symptoms that differ in manifestation or prevalence across the sexes – when investigating sex 'normal' differences because sex differences in ASC traits could mask such differences. However, this result should also be replicated using well-validated behavioural paradigms of local processing bias. It may be the case that males and females of the same ASC level do not differ in local versus global processing bias but simply interpret and respond to this particular item differently: a hypothesis that could be explored in future research by interviewing male and female respondents completing the AQ-10.

AQ32, referring to multi-tasking ability showed the opposite pattern of DIF. It was: 1) slightly less discriminating in males and 2) more likely to be endorsed by males given ASC level. We would argue that this most likely reflects normative sex differences that appear when ASC levels are taken into account. There is some evidence that females outperform males on multi-tasking paradigms (Stoet et al., 2013). The observed DIF effect implies that this previously identified female advantage is not simply a result of females tending have lower levels of ASC traits but a sex difference that exists independent of ASC. However, it is important to take into account the lay perception that multi-tasking is a more 'female' trait (e.g. see Stoet et al., 2013). This could affect the way that males and females respond to this item over and above their true multi-tasking ability and, for example, lead males to under-

report their multi-tasking ability because of a disinclination to endorse a female-typical behaviour.

Limitations and Future Directions

It is important to consider the potential limitations of the current study. First, although our sample size was large and included a broad range of ASC trait levels, including individuals with and without a clinical diagnosis, it was not a random draw from the population and can, therefore, not be considered population representative. We also had limited information about co-morbid psychopathology and could not, therefore, take this into account in our analyses. Future research should also examine the impact of co-morbidities especially on item responding. For example, ASC shows both overlap and co-morbidity with attention-deficit hyperactivity disorder (ADHD; Mayes, Calhoun, Mayes & Molitoris, 2012). As a result, there may be at least some items of the AQ-10 endorsed by respondents because of an underlying ADHD rather than ASC difficulty (e.g. see Sizoo et al. 2009).

It is also necessary to consider the possible impact that administering the AQ-10 in the context of the AQ-50 could have had on results. Previous studies have suggested that item responses are affected by the context in which they are administered (Desai & Braitman, 2005) and are e.g. primed by immediately preceding items (Weinberger et al., 2006) resulting in responses that are more similar across items presented close together than far apart (e.g. Harrison, McLaughlin & Coalter, 1996). While there is no reason to think that these kinds of effects should introduce or mask DIF or DTF by sex, this should be confirmed in future research administering the AQ-10 in isolation.

In terms of our statistical models, we also used a unidimensional IRT model, even though strict unidimensionality rarely holds in real data. Although unidimensional model fit well and PA-PCA, MAP and scree plots generally supported unidimensionality, these

methods do not necessarily indicate whether and to what degree parameter estimates were biased by any violations of the assumption (e.g. Bonifay, Reise, Scheines & Meijer; Reise, Scheines, Widaman & Haviland, 2013). Fortunately, IRT parameter estimates appear to be relatively robust to minor violations of unidimensionality (Kirisci, Hsu & Yu, 2001). Furthermore, the AQ-10 is used 'as if' unidimensional in practice, making it important to evaluate bias is in the test score in a manner corresponding to the way it is used to screen for ASC. In addition, we also used only two items as anchors to provide a common scale for the male and female IRT parameters; ideally, a larger number of non-DIF items would have been available for this purpose.

Finally, this study only addressed one source of diagnostic bias: that arising in screening for ASC prior to full diagnostic assessment. Interpretational biases by frontline clinicians and referrers are also likely to play a role. For example, anecdotal accounts suggest that social difficulties in females with ASC are more likely to be attributed to the person being 'just shy' (Lai, Lombardo, Auyeung, Chakrabarti & Baron-Cohen, 2015) or overlooked because of a focus on other symptoms (see Luciano, Keller, Politi, Aguglia & Magnano, 2014 for case study examples). Future research will be required to investigate other sources of diagnostic bias.

Conclusions

Although the AQ-10 contains some items that are sex-biased, these biases cancel out to give an overall unbiased test. The continued use of the AQ-10 as a brief screen for ASC is, therefore, supported. In addition, the nature of the biases in the items of the AQ-10 reveal possible differences in how specific ASC behaviours may differ by sex.

References

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*, 202-212.

American Psychiatric Association. (2013). *DSM 5*. American Psychiatric Association.

American Psychiatric Association. (2000). *DSM-IV-TR: Diagnostic and statistical manual of mental disorders, text revision*. American Psychiatric Association.

Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., & Charman, T. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *The Lancet*, *368*, 210-215.

Bauman, M. L. (2010). Medical comorbidities in autism: challenges to diagnosis and treatment. *Neurotherapeutics*, *7*, 320-327.

Baron-Cohen, S., Lombardo, M. V., Auyeung, B., Ashwin, E., Chakrabarti, B., & Knickmeyer, R. (2011). Why are autism spectrum conditions more prevalent in males?.*PLoS biology*, *9*, e1001081.

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., & Brayne, C. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *British Journal of Psychiatry*, *194*, 500-509.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-

functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*, 5-17.

Begeer, S., Mandell, D., Wijnker-Holmes, B., Venderbosch, S., Rem, D., Stekelenburg, F., & Koot, H. M. (2013). Sex differences in the timing of identification among children and adults with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 43*, 1151-1156.

Booth, T., Murray, A. L., McKenzie, K., Kuenssberg, R., O'Donnell, M., & Burnett, H. (2013). Brief report: An evaluation of the AQ-10 as a brief screening instrument for ASD in adults. *Journal of Autism and Developmental Disorders,43*, 2997-300.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index.*Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 504-516.

Buck, T. R., Viskochil, J., Farley, M., Coon, H., McMahon, W. M., Morgan, J., & Bilder, D. A. (2014). Psychiatric comorbidity and medication use in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders,44*, 3063-3071.

Carter, A. S., Black, D. O., Tewani, S., Connolly, C. E., Kadlec, M. B., & Tager-Flusberg, H. (2007). Sex differences in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*, 86-97.

Cassidy, S., Bradley, P., Robinson, J., Allison, C., McHugh, M., & Baron-Cohen, S. (2014). Suicidal ideation and suicide plans or attempts in adults with Asperger's syndrome attending a specialist diagnostic clinic: a clinical cohort study. *The Lancet Psychiatry, 1*, 142-147.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1-29

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76,* 114-140.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37-45.

Dworzynski, K., Ronald, A., Bolton, P., & Happé, F. (2012). How different are girls and boys above and below the diagnostic threshold for autism spectrum disorders?. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51,* 788-797.

Eaves, L. C., & Ho, H. H. (2008). Young adult outcome of autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *38*, 739-747.

Fombonne, E. (2009). Epidemiology of pervasive developmental disorders. *Pediatric Research*, *65*, 591-598.

Giarelli, E., Wiggins, L. D., Rice, C. E., Levy, S. E., Kirby, R. S., Pinto-Martin, J., & Mandell, D. (2010). Sex differences in the evaluation and diagnosis of autism spectrum disorders among children. *Disability and Health Journal*, *3*, 107-116.

Golding, J. (2011). Social and demographic factors that influence the diagnosis of autistic spectrum disorders. *Social Psychiatry and Psychiatric Epidemiology*, *46*, 1283-1293.

Halladay, A. K., Bishop, S., Constantino, J. N., Daniels, A. M., Koenig, K., Palmer, K., ... & Szatmari, P. (2015). Sex and gender differences in autism spectrum disorder:

summarizing evidence gaps and identifying emerging areas of priority. *Molecular Autism, 6,* 36.

Hartley, S. L., & Sikora, D. M. (2009). Sex differences in autism spectrum disorder: An examination of developmental functioning, autistic symptoms, and coexisting behavior problems in toddlers. *Journal of Autism and Developmental Disorders, 39,* 1715-1722.

Hoekstra, R. A., Vinkhuyzen, A. A., Wheelwright, S., Bartels, M., Boomsma, D. I., Baron-Cohen, S., ... & van der Sluis, S. (2011). The construction and validation of an abridged version of the autism-spectrum quotient (AQ-Short).*Journal of Autism and Developmental Disorders, 41,* 589-596.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1-55.

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8,* 291-312.

Kim, Y. S., Leventhal, B. L., Koh, Y. J., Fombonne, E., Laska, E., Lim, E. C., ... & Grinker, R. R. (2011). Prevalence of autism spectrum disorders in a total population sample. *American Journal of Psychiatry, 168,* 904-912.

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied psychological measurement, 25,* 146-162.

Kopp, S., & Gillberg, C. (2011). The Autism Spectrum Screening Questionnaire (ASSQ)-Revised Extended Version (ASSQ-REV): An instrument for better capturing the autism phenotype in girls? A preliminary study involving 191 clinical cases and community controls. *Research in Developmental Disabilities,32*, 2875-2888.

Kreiser, N. L., & White, S. W. (2014). ASC in females: Are we overstating the gender difference in diagnosis?. *Clinical Child and Family Psychology Review,17*, 67-84.

Lai, M. C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/Gender Differences and Autism: Setting the Scene for Future Research. *Journal of the American Academy of Child & Adolescent Psychiatry,54*, 11-24.

Lai, M. C., Lombardo, M. V., Pasco, G., Ruigrok, A. N., Wheelwright, S. J., Sadek, S. A., ... & MRC AIMS Consortium. (2011). A behavioral comparison of male and female adults with high functioning autism spectrum conditions. *PloS one*, *6*, e20835.

Lai, M. C., Lombardo, M. V., Suckling, J., Ruigrok, A. N., Chakrabarti, B., Ecker, C., ... & Baron-Cohen, S. (2013). Biological sex affects the neurobiology of autism. *Brain*, *136*, 2799-2815.

Luciano, C. C., Keller, R., Politi, P., Aguglia, E., & Magnano, F. (2014). Misdiagnosis of high function autism spectrum disorders in adults: An Italian case series. *Autism*, *4*, 2.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847-862.

Matson, J. L., & Shoemaker, M. (2009). Intellectual disability and its relationship to autism spectrum disorders. *Research in Developmental Disabilities*, *30*, 1107-1114.

Mayes, S. D., Calhoun, S. L., Mayes, R. D., & Molitoris, S. (2012). Autism and ADHD: Overlapping and discriminating symptoms. *Research in Autism Spectrum Disorders, 6,* 277-285.

National Institute for Health and Clinical Excellence (NICE) editor. Autism: recognition, referral, diagnosis and management of adults on the autism spectrum. Clinical Guideline 142 ed. London: National Institute for Health and Clinical Excellence; 2012.

Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2015). What Range of Trait Levels Can the Autism-Spectrum Quotient (AQ) Measure Reliably? An Item Response Theory Analysis. in press *Psychological Assessment.*

Murray, A. L., Booth, T., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are Autistic traits measured equivalently in individuals with and without an Autism Spectrum Disorder? An invariance analysis of the Autism Spectrum Quotient Short Form. *Journal of Autism and Developmental Disorders, 44,* 55-64.

Murray, A. L., Kuenssberg, R., McKenzie, K. & Booth, T. (2015). Do the Autism Spectrum Quotient (AQ) and Autism Spectrum Quotient Short Form (AQ-S) primarily reflect general ASD traits or specific ASD traits?: A bi-factor analysis. in press *Assessment.*

Murray, A. L., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are we under-estimating the association between autism symptoms?: The importance of considering simultaneous selection when using samples of individuals who meet diagnostic criteria for an autism spectrum disorder. *Journal of Autism and Developmental Disorders,* 1-10.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational and Behavioral Statistics, 7*, 105-118.

Muthén, L., & Muthén, B. (2010). Mplus 6.11 version user's guide. *Los Angeles, CA.*

Pletzer, B. (2014). Sex-specific strategy use and global-local processing: a perspective toward integrating sex differences in cognition. *Frontiers in Neuroscience, 8*, 425.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-164.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. *Educational and Psychological Measurement, 73*, 5-26.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*, 1-25.Russell, G., Steer, C., &

Rutherford, M., McKenzie, K., Johnson, T., Catchpole, C., O'Hare, A., McClure, I., Forsyth, K., McCartney, D., Murray, A. L. (2015). Gender ratio in a clinical population sample, age of diagnosis and the duration of assessment procedures in children and adults with Autism Spectrum Disorder. Submitted.

Rutter, M., Caspi, A., & Moffitt, T. E. (2003). Using sex differences in psychopathology to study causal mechanisms: unifying issues and research strategies. *Journal of Child Psychology and Psychiatry, 44*, 1092-1115.

Sass, D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 347-363.

Shattuck, P. T., Durkin, M., Maenner, M., Newschaffer, C., Mandell, D. S., Wiggins, L., ... & Cuniff, C. (2009). Timing of identification among children with an autism spectrum disorder: findings from a population-based surveillance study. *Journal of the American Academy of Child & Adolescent Psychiatry*,*48*, 474-483.

Sizoo, B. B., van den Brink, W., Gorissen-van Eenige, M., Koeter, M. W., van Wijngaarden-Cremers, P. J., & van der Gaag, R. J. (2009). Using the Autism-Spectrum Quotient to discriminate Autism Spectrum Disorder from ADHD in adult patients with and without comorbid substance use disorder. *Journal of Autism and Developmental Disorders*, *39*, 1291-1297.

Stoet, G., O'Connor, D. B., Conner, M., & Laws, K. R. (2013). Are women better than men at multi-tasking?. *BMC Psychology*, *1*, 18.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Unpublished doctoral dissertation, University of California, Los Angeles.

Zwaigenbaum, L., Bryson, S. E., Szatmari, P., Brian, J., Smith, I. M., Roberts, W., ... & Roncadin, C. (2012). Sex differences in children with autism spectrum disorder identified within a high-risk infant cohort. *Journal of Autism and Developmental]Disorders*, *42*, 2585-2596.

Table 1

Item endorsement for male and female subsamples

Item	Item Endorsement	
	Males	**Females**
I often notice small sounds when others do not (5).	.65	.63
I usually concentrate more on the whole picture, rather than the small details (28).	.44	.48
I find it easy to do more than one thing at once (32).	.45	.26
If there is an interruption, I can switch back to what I was doing very quickly (37).	.44	.40
I find it easy to 'read between the lines' when someone is talking to me (27)	.38	.27
I know how to tell if someone listening to me is getting bored (31)	.28	.19
When I'm reading a story I find it difficult to work out the characters' intentions (20)	.25	.17
I like to collect information about categories of things (e.g., types of car, types of bird, types of train, types of plant, etc.) (41)	.43	.28
I find it easy to work out what someone is thinking or feeling just by looking at their face (36)	.38	.24
I find it difficult to work out people's intentions (45)	.41	.32
AQ-10 total score	4.1 (SD=2.8)	3.2 (SD=2.6)

Note. Items are coded such that endorsing an item represents higher levels of autistic traits. AQ item numbers in parentheses

Table 2

Single group CFA model fits

Group	χ^2	df	*P*	CFI	TLI	RMSEA	WRMR
Males	462.35	35	<.001	.97	.96	.06	2.34
Females	580.84	35	<.001	.97	.96	.06	2.59

Table 3

2PL model parameter estimates and DIF analysis of AQ-10 items

Item	Content	Male a	Male b	Female a	Female b	χ^2	*p*	ΔAIC	ΔsaBIC	ΔBIC
5	I often notice small sounds when others do not.	0.39	0.69	0.39	0.69	-	-	-	-	-
28	I usually concentrate more on the whole picture, rather than the small details. R	0.92	-0.26	0.77	0.31	96.54	<.01	92.54	84.78	78.42
32	I find it easy to do more than one thing at once. R	1.16	-0.24	1.35	-0.81	48.27	<.01	44.27	36.51	30.15
37	If there is an interruption, I can switch back to what I was doing very quickly. R	1.06	-0.31	1.09	0.03	27.84	<.01	23.83	16.07	9.72
27	I find it easy to 'read between the lines' when someone is talking to me. R	2.73	-0.92	2.44	-0.73	3.88	.14	-0.12	-7.88	-14.23
31	I know how to tell if someone listening to me is getting bored. R	1.89	-1.49	1.73	-1.44	1.66	.44	-2.34	-10.10	-16.46
20	When I'm reading a story I find it difficult to work out the characters' intentions	1.57	-1.56	1.59	-1.58		-	-	-	-
41	I like to collect information about categories of things (e.g., types of car, types of bird, types of train, types of plant, etc.).	0.87	-0.34	1.00	-0.71	28.34	<.01	24.34	2.23	2.23
36	I find it easy to work out what someone is thinking or feeling just by looking at their face. R	2.67	-0.94	2.46	-1.01	1.79	.41	-24.34	-2.23	-2.23
45	I find it difficult to work out people's intentions.	2.92	-0.67	2.19	-0.28	19.55	<.01	48.69	4.45	4.45

Note. R denotes reverse scored. Negative values of ΔAIC, ΔsaBIC and ΔBIC suggest that the model with no invariance constraints is better fitting.

Figure 1

Example of a hypothetical item showing uniform DIF

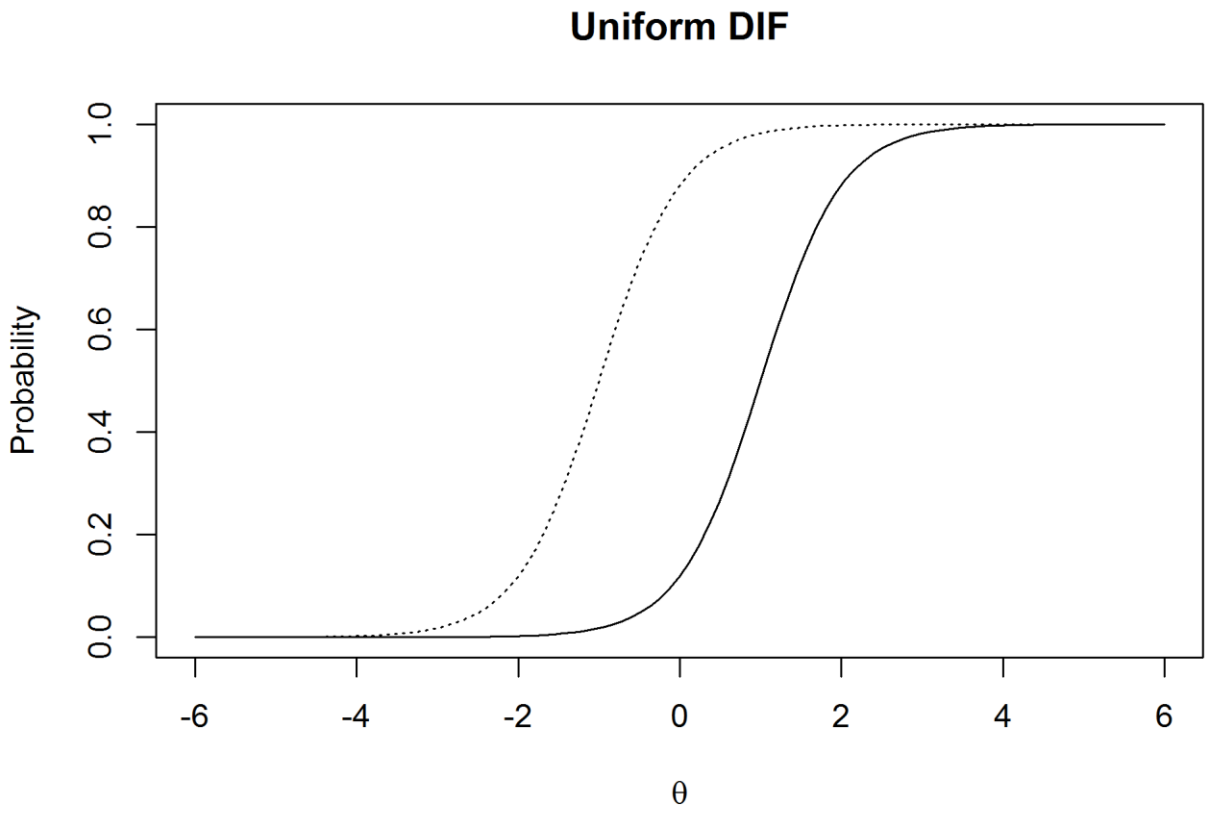

Figure 2

Example of a hypothetical item showing non-uniform DIF

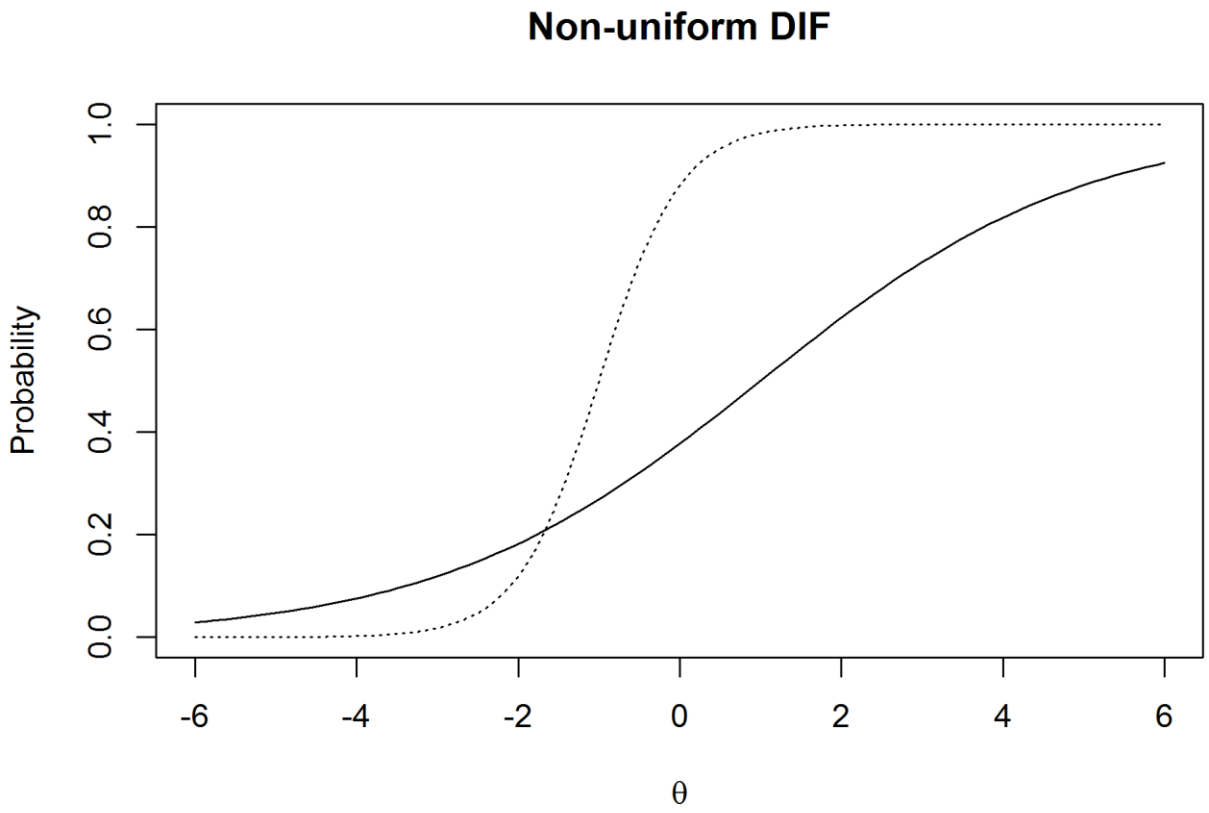

Figure 3

Item characteristic curves for males and females for AQ28

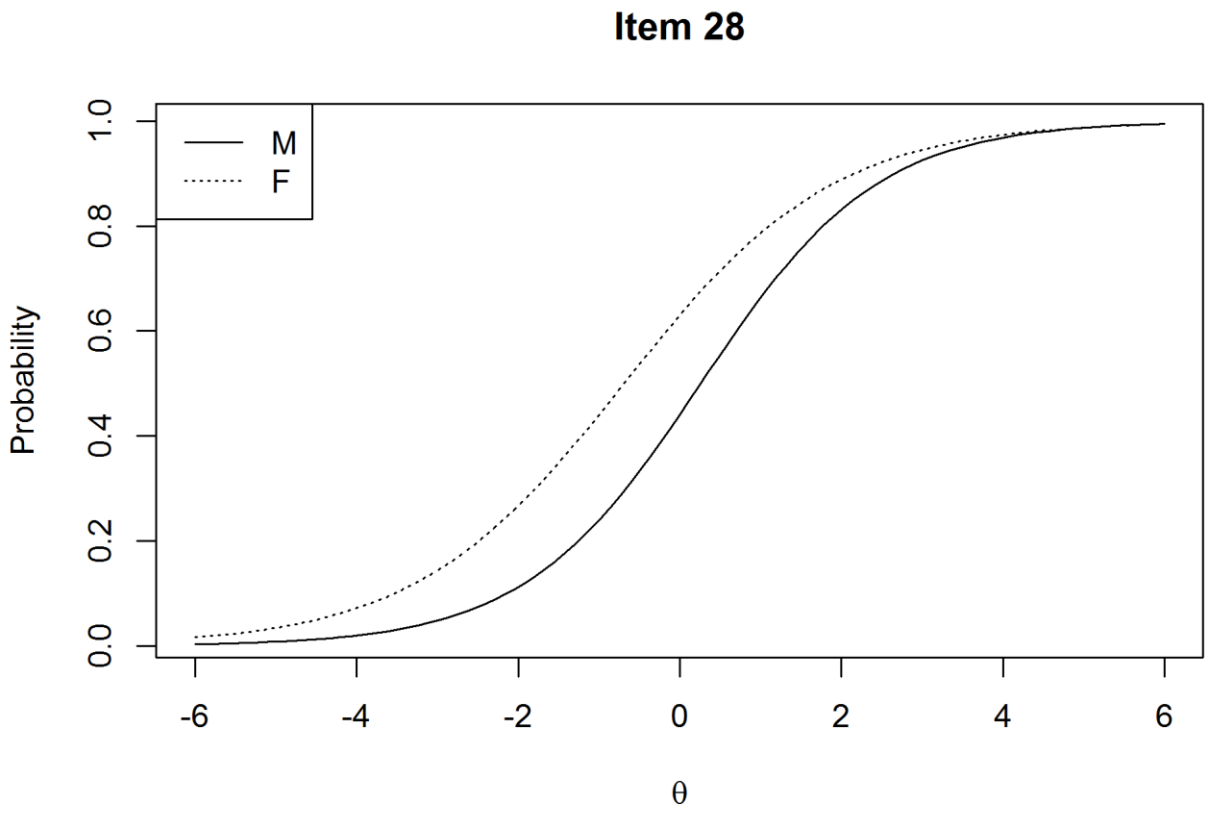

Figure 4

Item characteristic curves for males and females for AQ32

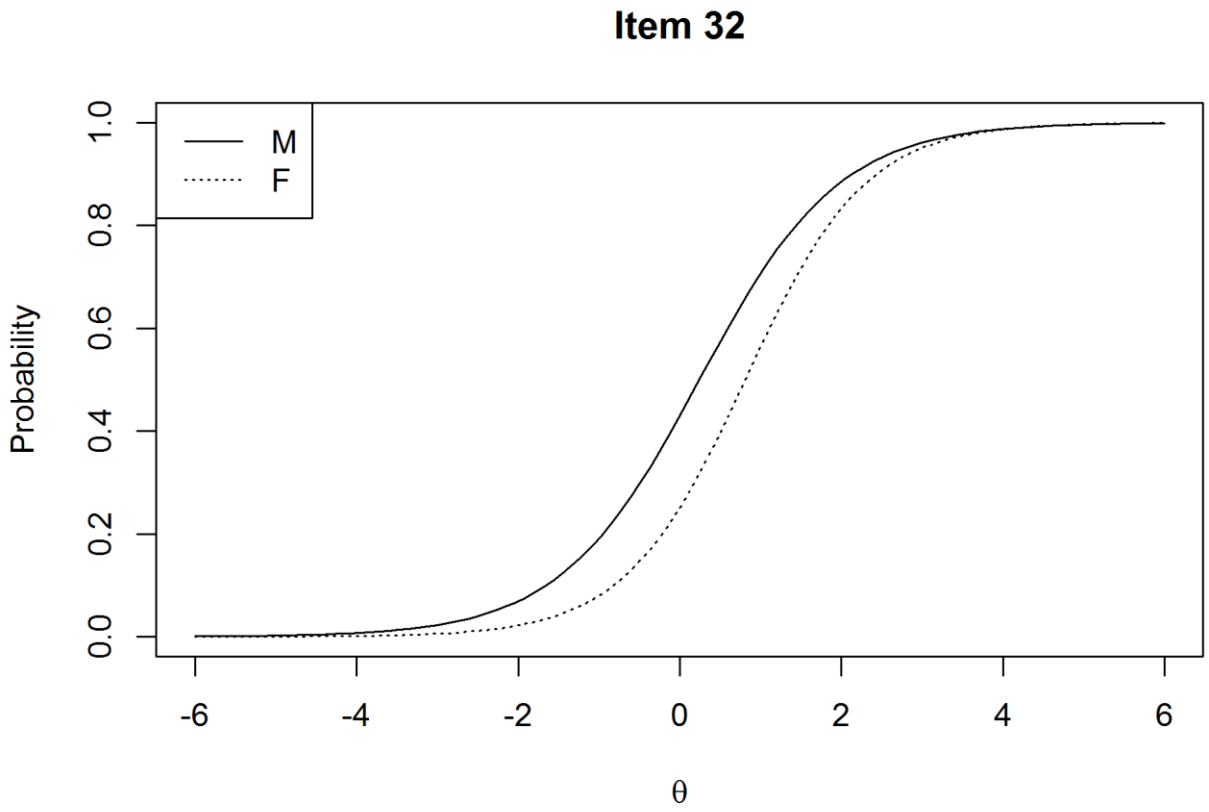

Figure 5

Test characteristic curves for males and females (AQ-10 total scores)